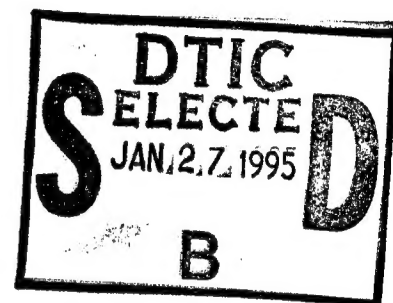


Segment-based Acoustic Models for Continuous Speech Recognition

Progress Report: 1 April 94 – 30 June 94

submitted to
Office of Naval Research
and
Advanced Research Projects Administration

by
Boston University
Boston, Massachusetts 02215



Principal Investigators

Dr. Mari Ostendorf
Associate Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Dr. J. Robin Rohlicek
Division Scientist, BBN Inc.
Telephone: (617) 873-3894

Administrative Contact

Maureen Rogers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365



DTIC QUALITY INSPECTED 3

19950124 003

Executive Summary

This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by extending previous work in segment-based modeling, by introducing a new hierarchical approach to representing intra-utterance statistical dependencies, and by developing language models that capture topic dependencies. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through a multi-pass search strategy that involves rescoring a constrained space given by an HMM decoding. We expect these different modeling techniques to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

With these overall project goals, the primary research efforts and results over the past quarter have included:

- experimentation with a new approach to continuous density parameter adaptation;
- improved the language modeling software to handle more general vocabularies and reduce storage requirements, and implemented a course-grained parallel training algorithm;
- extended the training algorithm for discrete distribution dependence trees (our model of intra-utterance correlation) to handle missing observations with the EM algorithm;
- implemented a dynamic programming algorithm for word lattice rescoring algorithm and demonstrated performance comparable to N-best rescoring with the SSM; and
- evaluated different channel compensation algorithms on the WSJ spoke S6 telephone recordings, achieving better results than those previously reported for this task.

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>See ADA 280332</i>	
Distribution	
Availability Codes	
Dist	Avail and/or Special
<i>A-1</i>	

Contents

1	Productivity Measures	4
2	Summary of Technical Progress	5
3	Publications and Presentations	11
4	Transitions and DoD Interactions	12
5	Software and Hardware Prototypes	13

Principal Investigator Name: Mari Ostendorf

PI Institution: Boston University

PI Phone Number: 617-353-5430

PI E-mail Address: mo@raven.bu.edu

Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition

Grant or Contract Number: ONR-N00014-92-J-1778

Reporting Period: 1 April 1994 – 30 June 1994

1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 0
- Unrefereed reports and articles: 1
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 2
- Contributed presentations: 1
- Honors received: none
- Prizes or awards received: none
- Promotions obtained: none
- Graduate students supported $\geq 25\%$ of full time: 5
- Post-docs supported $\geq 25\%$ of full time: 0
- Minorities supported: 2 women

Principal Investigator Name: Mari Ostendorf

PI Institution: Boston University

PI Phone Number: 617-353-5430

PI E-mail Address: mo@raven.bu.edu

Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition

Grant or Contract Number: ONR-N00014-92-J-1778

Reporting Period: 1 April 1994 – 30 June 1994

2 Summary of Technical Progress

Introduction and Background

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and primarily in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels: the frame level (e.g. signal processing), the phoneme level (e.g. modeling feature dynamics), and the utterance level (e.g. defining a structural context for representing the intra-utterance dependence across phonemes). This project addresses the problem of acoustic modeling, specifically focusing on modeling at the segment level and above. The research strategy includes three main thrusts. First, phone-level acoustic modeling is based on the stochastic segment model (SSM) [1, 2], and in this area our main efforts involve developing new techniques for robust context modeling, mechanisms for effectively incorporating segmental features, and models of within-segment dependence of frame-based features. Second, high-level models are being explored in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model. In particular, we are investigating hierarchical structures for representing the intra-utterance dependency of phonetic models, and more recently language models for representing topic dependency and language dynamics, recognizing that higher-order models of correlation can extend to the language domain as well as the acoustic domain. Lastly, speech recognition is implemented under a multi-pass search framework, which in most of our work has been based on the N-best rescoring paradigm [3] where we use the BBN Byblos system is used to constrain the SSM search space by providing the top N sentence hypotheses. This paradigm facilitates research on high-order models through reducing development costs, and provides a modular framework for technology transfer that has enabled us to advance state-of-the-art recognition performance through collaboration with BBN.

Summary of Technical Results

In brief, the accomplishments of previous work on this project have included: improvements to the N-Best rescoring weight estimation algorithm; investigation of different mechanisms for improving the baseline acoustic model, including distribution clustering [4], mixture modeling at different time scales [5, 6], theoretically consistent models based on context-dependent posterior distributions, automatic distribution mapping estimation, and hierarchical models of intra-utterance phoneme dependence; development of a new approach to adaptation of continuous density parameters; implementation of baseline n-gram and sentence-level mixture language models [10]; and improvements to the SSM baseline system aimed at improving performance on the ARPA WSJ task and participation in the benchmark tests.

The research efforts during this quarter, supported in part by AASERT awards, have primarily involved theoretical and initial software development for models of high-order correlation and search algorithms to accommodate these models. These efforts and other related research developments are summarized below.

Adaptation of continuous distributions. It is well known that speech recognition performance can be improved by using automatic adaptation techniques to tune a general model to a particular condition, whether it be to address channel/environment effects or match the recognizer to a particular speaker. In particular, we are interested in incremental adaptation of Gaussian density parameters, which is more difficult than adaptation of tied mixture parameters (the approach currently used by most sites) because of the large number of parameters to adapt. In the last quarter, we developed an approach for adaptation that combines the advantages of a Bayesian framework [7, 8] with the robust properties of vector field smoothing [9] in a two-level model built using divisive, maximum likelihood distribution clustering. Adaptation of the detailed level is based on a weighted combination of statistics accumulated at a coarse level.

In this quarter, we implemented the algorithm and began experimentation in supervised, incremental adaptation mode on the WSJ H2 task. We compared two methods for computing the weights used in adaptation: one based on Euclidean distance similar to that used in vector field smoothing and one based on an information-theoretic similarity measure. Preliminary results show that the information theoretic measure gives better performance than the Euclidean distance measure. However, neither method gave as large a gain as is reported in other work for vector field smoothing. We suspect that this is because not all model correlations were used to reduce computation and storage costs in the initial implementation for finding the mutual information between different distributions. Therefore, future work will involve software design efforts to reduce the cost of training adaptation coefficients.

Mixture language modeling. One of the important questions in language modeling today is how to effectively represent the long-term structure of language, i.e. how to capture dependence over longer sequences of words than can be modeled with a simple n-gram. To address this problem, we have developed a sentence-level mixture language model (LM) that represents the topic-dependent structure of language with separate n-gram language model mixture components determined using automatic clustering. In the previous quarter, we obtained a 6% reduction in recognition error using a 5-component mixture models as compared to the standard trigram models on the 5k vocabulary WSJ H2 task.

In this quarter, we focused on software development to handle large vocabularies. As the language model training data and the vocabulary size increase, storing the statistics of the language model becomes an important issue in using the language model for recognition. Currently, with 72 million words of training data and 5K vocabulary, there are approximately 14 million trigrams and 2.2 million bigrams present. In order to store all these statistics in an easily accessible manner and at the same time reduce the storage requirement, we conducted a few experiments for model size reduction. We observed that storing the probabilities to a lesser precision, by using a log-base representation or by considering the probability values to their third decimal place, did not lead to a significant increase in perplexity. However, it reduced storage requirements by a significant 30%. We have also developed code for coarse-grained parallel training of and scoring with the language model to reduce computation time. In addition, we have made the code more general to handle increases in both vocabulary size and language model training data.

Current efforts are aimed at implementing more theoretically motivated algorithms, such as a similarity measure using inverse document frequencies and full Expectation-Maximization (EM) training, which we expect will result in a small improvement in the models.

Intra-utterance phoneme dependence modeling. We further developed the theoretical framework for a hierarchical model of dependence for a set of discrete random variables, which we have begun investigating as a model of intra-utterance phoneme dependence. We use a dependence tree [11] to represent the correlation among random variables, using a tree structure (designed automatically) with Markov assumptions along the branches of the tree. The dependence tree gives us an approximation to the joint distribution of the set of random variables with a much smaller number of parameters than a full joint model, and thus a feasible way of modeling intra-utterance correlation of phones in speech recognition applications. The dependence tree can be thought of as representing a vector "state" that describes the speaker/utterance, where each element of the vector corresponds to a phoneme. Since most utterances will not contain all possible phonemes, we derived an efficient algorithm for computing the likelihood of the observed data, which we call the upward-downward algorithm to emphasize the analogy to the forward-backward algorithm. This algorithm is needed for solving the parameter estimation problem when the tree structure is given, which is then used as one step in an iterative approach to combined dependence tree topology

design and parameter estimation. We have started to implement the EM algorithm for training the model and plan to combine this with the tree topology design algorithm and run initial experiments on the TIMIT corpus.

Lattice search algorithms for multi-pass recognition scoring. In previous work, we implemented an SSM word lattice rescoring algorithm that uses a dynamic programming (DP) search and handles cross-word triphone models and trigram language models. In this quarter, we modified the BBN decoder to produce lattices from the N-best traceback structure that are annotated with HMM segmentations and score information. The modifications to the BBN decoder were implemented by BU student Fred Richardson at BBN, so the software can be used in other BBN work. In addition, we ran experiments that verified that performance of the DP search with the SSM is comparable to N-best rescoring even though the time information is based on a less detailed HMM. In future work, we plan to implement and assess trade-offs of a local search algorithm for rescoring the lattice to handle long-distance knowledge sources. [This work was supported by an ONR AASERT award.]

Channel Modeling Two existing methods for channel compensation were evaluated as baselines for further work. We compared cepstral mean subtraction (CMS) to the maximum likelihood (ML) channel estimate introduced in [12]. In both cases, the channel estimate is subtracted from the cepstral vectors of the speech before the BU recognition pass. For CMS, the estimate is simply the average of the cepstral vectors for the utterance, whereas the estimate is compensated by the hypothesized phone sequence in the ML approach. Our implementation of the ML estimate is not strictly ML, since we use the HMM top hypothesized phone-sequences rather than multiple EM iterations, but we do not think that performance will be affected and it is computationally efficient. Experiments were conducted on the Wall Street Journal 1993 S6 telephone data task. Like Neumeyer *et al.* [12], we found no benefit to the more expensive ML approach, as indicated in Table 1. However, we did achieve word error rates in both cases that were lower than those reported by all other sites on this task: 11.2% vs. rates of 12.5-25.3% reported at the March 1994 ARPA SLT workshop. Despite the somewhat discouraging results for the ML algorithm, we plan to continue to explore more sophisticated channel estimation algorithms, because we believe that the differences in performance will be significant for short utterances. [This work was supported by an ARPA AASERT award associated with this project.]

Future Goals

Based on the results of the past year and our original goals for the project, we have set the following goals for the next six months: (1) improve the implementation of adaptation training and further experiment with incremental adaptation; (2) extend the language modeling work to

Table 1: Word error rate for the SSM recognition system on the WSJ S6 telephone data, training on a band-limited version of the WSJ1 training set.

WSJ Test	Channel Estimation	Word Error Rate	
		Dev	Eval
S6	CMS	11.6	11.2
S6	ML	11.6	11.5

include a dynamic component; (3) implement the hierarchical model training algorithm and run dependence tree design experiments on the TIMIT corpus; (4) implement the lattice local search algorithm and assess performance/speed trade-offs of the different lattice search algorithms; and (5) develop a new approach to channel estimation.

References

- [1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing*, Dec. 1989.
- [2] S. Roukos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 127-130, New York, New York, April 1988.
- [3] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. of the DARPA Workshop on Speech and Natural Language*, pp. 83-87, February 1991.
- [4] A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, to appear.
- [5] A. Kannan and M. Ostendorf, "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing*, Vol. II, pp. 327-330, April 1993.
- [6] O. Kimball and M. Ostendorf, "On the Use of Tied Mixture Distributions," *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 102-107, 1993.
- [7] C.H. Lee, C.H. Lin, and B.H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans on Signal Processing*, Vol. 39, No. 4, April 1991, pp. 806-814.

- [8] B. Necioglu, M. Ostendorf and J. R. Rohlicek, "A Bayesian Approach to Speaker Adaptation of the Stochastic Segment Model," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing*, March 1992, Vol. I, pp. 437-440.
- [9] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," *Proc. of the Inter. Conf. on Spoken Language Processing*, Vol. 1, pp. 369-372.
- [10] R. Iyer, M. Ostendorf and J. R. Rohlicek, "Language Modeling with Sentence-Level Mixtures," *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, March 1994.
- [11] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, Vol. IT-14, No. 3, May 1968, pp. 462-467.
- [12] L. Newmeyer, V. Digalakis, M. Weintraub, "Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus," *IEEE Transactions on Speech and Audio Processing*, 1994.

Principal Investigator Name: Mari Ostendorf

PI Institution: Boston University

PI Phone Number: 617-353-5430

PI E-mail Address: mo@raven.bu.edu

Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition

Grant or Contract Number: ONR-N00014-92-J-1778

Reporting Period: 1 April 1994 – 30 June 1994

3 Publications and Presentations

During this reporting period, we published one conference paper, gave two invited talks and one conference presentation, as itemized below. In addition, one Boston University M.S. thesis proposal was successfully defended by Rukmini Iyer, entitled "Language Modeling with Sentence-Level Mixtures."

Unrefereed reports and articles:

"Stochastic Segment Modeling for CSR: the BU WSJ Benchmark System," M. Ostendorf, F. Richardson, S. Tibrewal, R. Iyer, O. Kimball, and J. R. Rohlicek, *Proceedings of the 1994 ARPA Workshop on Spoken Language Technology*.

Conference presentations and invited talks:

"A Unified View of Stochastic Modeling for Speech Recognition", M. Ostendorf, invited talk at DRA in Malvern, UK, and at Dragon Systems in Massachusetts.

"Adaptation of Continuous Gaussian Densities," M. Ostendorf, presented at the June 1994 *Speech Research Symposium*.

Principal Investigator Name: Mari Ostendorf

PI Institution: Boston University

PI Phone Number: 617-353-5430

PI E-mail Address: mo@raven.bu.edu

Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition

Grant or Contract Number: ONR-N00014-92-J-1778

Reporting Period: 1 April 1994 - 30 June 1994

4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes.

We have also begun an effort to collaborate more closely in lattice rescoring. Boston University student Fred Richardson has implemented software libraries that will be shared by both sites, and he has modified the BBN decoder to provide lattices annotated with segmentation times and HMM scores.

The recognition system that has been developed under the support of this grant and of a joint NSF-ARPA grant (NSF # IRI-8902124) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University in a project supported by the Linguistic Data Consortium.

Principal Investigator Name: Mari Ostendorf

PI Institution: Boston University

PI Phone Number: 617-353-5430

PI E-mail Address: mo@raven.bu.edu

Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition

Grant or Contract Number: ONR-N00014-92-J-1778

Reporting Period: 1 April 1994 – 30 June 1994

5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.